IJARSCT
Impact Factor: 7.67
IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Volume 5, Issue 8, March 2025

Anti-Cancer Drug Response Prediction using Gene Expression

Prof. Alka Kumbhar[1], Mr. Tushar Wankhade[2], Mr. Rushikesh Shinde[3],
Mr. Prathamesh Pawar[4],Mr. Gaurav Naiknaware[5]
Prof., Computer Engineering, Genba Sopanrao Moze College of Engineering, Balewadi, Pune, India [1]
Student, Computer Engineering, Genba Sopanrao Moze College of Engineering, Balewadi, Pune, India[2,3,4,5]

Abstract*: This project aims to develop and evaluate machine learning models for predicting the response of cancer cells to anti-cancer drugs, utilizing gene expression data. The project leverages publicly available datasets like the Genomics of Drug Sensitivity in Cancer (GDSC) and the Cancer Cell Line Encyclopedia (CCLE), which contain gene expression profiles and drug response measurements (e.g., IC50, AUC) for various cancer cell lines. By training machine learning algorithms on this data, the project seeks to build predictive models that can accurately identify effective drug treatments for individual cancer cell lines, contributing to the advancement of precision oncology.They emphasize the need to tailor cancer treatments to individual patients, driven by the variability in drug response due to tumor heterogeneity.Abstracts frequently mention the utilization of large-scale datasets like the Genomics of Drug Sensitivity in Cancer (GDSC) and the Cancer Cell Line Encyclopedia (CCLE). These datasets provide valuable gene expression and drug response information.*

Keywords: Drug Response Prediction, Predictor Genes, Cell Line Gene Expression Data, Machine Learning (ML)

I. INTRODUCTION

Although many models have been proposed to accurately predict the response of drugs in cell lines recent years, understanding the genome related to drug response is also the key for completing oncology precision medicine. In this paper, based on the cancer cell line gene expression and the drug response data, we established a reliable and accurate drug response prediction model and found predictor genes for some drugs of interest. To this end, we first performed pre-selection of genes based on the Pearson correlation coefficient and then used Elastic Net regression model for drug response prediction and fine gene se lection. To find more reliable set of predictor genes, we performed regression twice for each drug, one with IC50 and the other with area under the curve (AUC) (or activity area). For the 12 drugs we tested, the predictive performance in terms of Pearson correlation coefficient exceeded 0.6 and the highest one was 17-AAG for which Pearson correlation coefficient was 0.811 for IC50 and 0.81 for AUC. We identify common predictor genes for IC50 and AUC, with which the performance was similar to those with genes separately found for IC50 and AUC, but with much smaller number of predictor genes. By using only common predictor genes, the highest performance was AZD6244 (0.8016 for IC50, 0.7945 for AUC) with 321 predictor genes.

Cancer is one of main causes of death worldwide. Anti-cancer drug therapy is an import ant part of cancer treatment and an effective use of them can prolong patient's survival. According to many clinical data, patients with the same cancer have quite different response to the same treatment or the same drugs due to genomic specificity. Recently, targeted anti-cancer therapy [1,2] considering gene-specific effects has been proposed as a new cancer therapy. In order to develop specific targeted therapy for cancer patients in clinical treatment, many clinical trials are required. However, there are many obstacles such as sample limitations, complicated operations, high environmental requirements, and high cost, which far from meeting the demand. With the rapid development of artificial intelligence, many machine

ISSN 2581-9429 IJARSCT

DOI: 10.48175/568

learning based drug response prediction models were proposed utilizing genomic information and anti-cancer drug response data.

In this report, we explore a machine learning-based approach for predicting anti-cancer drug response using gene expression data from cancer cell lines. Using the Genomics of Drug Sensitivity in Cancer (GDSC) database, we focus on 6-7 drugs and their effects on more than 5,000 cancer cell lines. By employing the ElasticNet regression model, we aim to predict drug response with high accuracy and identify predictor genes that are consistently associated with drug sensitivity. This report also delves into the biological significance of the identified genes through enrichment analysis, linking these predictor genes to critical cancer-related pathways. Ultimately, our goal is to enhance the predictive power of machine learning models in cancer therapy and pave the way for more personalized and effective treatments.

Project Objectives:

Data Acquisition and Preprocessing:

- Download and integrate gene expression and drug response data from GDSC and CCLE.
- Perform data cleaning, normalization, and feature scaling to prepare the data for machine learning.
- Address missing values and handle potential batch effects.

Feature Selection and Dimensionality Reduction:

- Implement feature selection techniques (e.g., Pearson correlation, Recursive Feature Elimination) to identify relevant genes.
- Explore dimensionality reduction methods (e.g., Principal Component Analysis, t-SNE) to reduce data complexity.

Model Development and Training:

- Implement and evaluate various machine learning algorithms, including regression models (Ridge, ElasticNet), Support Vector Machines (SVM), and potentially deep learning neural networks.
- Optimize model hyperparameters using cross-validation techniques.

Model Evaluation and Validation:

- Evaluate model performance using appropriate metrics (e.g., Pearson correlation coefficient, Root Mean Squared Error, AUC).
- Perform rigorous validation using independent test sets to assess model generalization.
- Compare the performance of the different machine learning models that are used.

Interpretation and Analysis:

- Analyze the identified relevant genes and their association with drug response.
- Explore the potential biological implications of the findings.
- Document all findings, and create a report.

II. LITERATURE REVIEW

Improved Anticancer Drug Response Prediction in Cell Lines Using Matrix Factorization with Similarity Regularization (2017)

Authors: Lin Wang, Xiaozhong Li, Louxin Zhang, Qiang Gao

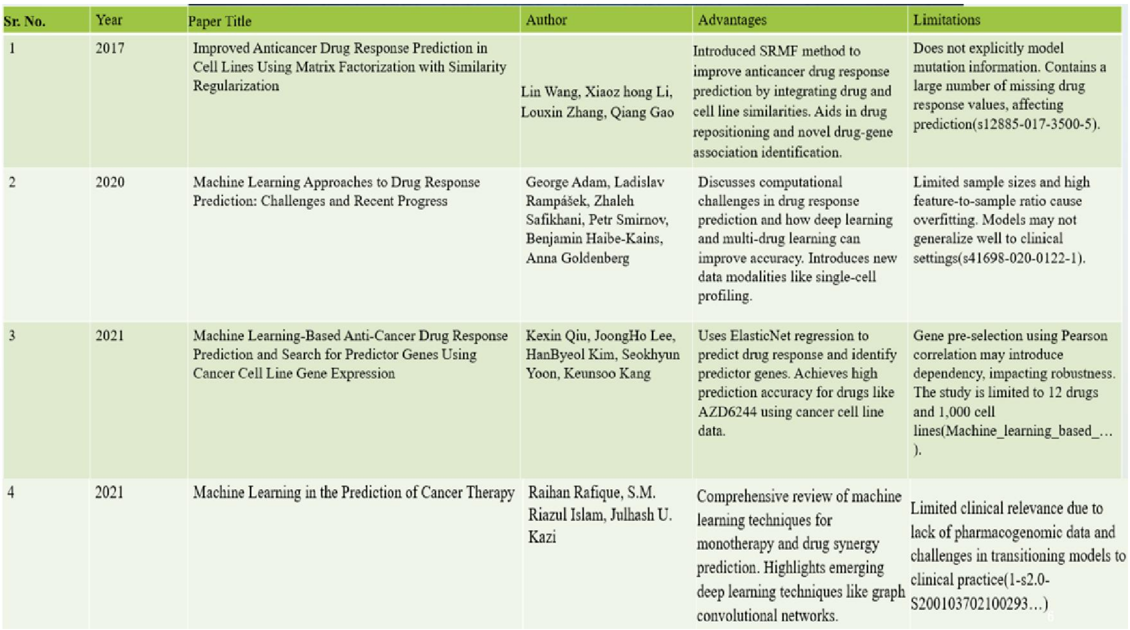
Description:

This study introduced the Similarity Regularized Matrix Factorization (SRMF) method to improve drug response predictions by integrating drug and cell line similarities. The approach aids in drug repositioning and helps identify novel drug-gene associations.

Advantages: SRMF enhances predictive performance by considering drug and cell line similarities, making it useful for drug repositioning and gene association studies.

ISSN 2581-9429 IJARSCT
DOI: 10.48175/568

Limitations: The model lacks explicit information about its integration process, and a significant amount of missing drug response data reduces prediction accuracy.



Sr. No.	Year	Paper Title	Author	Advantages	Limitations
1	2017	Improved Anticancer Drug Response Prediction in Cell Lines Using Matrix Factorization with Similarity Regularization	Lin Wang, Xiaoz hong Li, Louxin Zhang, Qiang Gao	Introduced SRMF method to improve anticancer drug response prediction by integrating drug and cell line similarities. Aids in drug repositioning and novel drug-gene association identification.	Does not explicitly model mutation information. Contains a large number of missing drug response values, affecting prediction(s12885-017-3500-5).
2	2020	Machine Learning Approaches to Drug Response Prediction: Challenges and Recent Progress	George Adam, Ladislav Rampášek, Zhaleh Safikhani, Petr Smirnov, Benjamin Haibe-Kains, Anna Goldenberg	Discusses computational challenges in drug response prediction and how deep learning and multi-drug learning can improve accuracy. Introduces new data modalities like single-cell profiling.	Limited sample sizes and high feature-to-sample ratio cause overfitting. Models may not generalize well to clinical settings(s41698-020-0122-1).
3	2021	Machine Learning-Based Anti-Cancer Drug Response Prediction and Search for Predictor Genes Using Cancer Cell Line Gene Expression	Kexin Qiu, JoongHo Lee, HanByeol Kim, Seokhyun Yoon, Keunsoo Kang	Uses ElasticNet regression to predict drug response and identify predictor genes. Achieves high prediction accuracy for drugs like AZD6244 using cancer cell line data.	Gene pre-selection using Pearson correlation may introduce dependency, impacting robustness. The study is limited to 12 drugs and 1,000 cell lines(Machine_learning_based_…).
4	2021	Machine Learning in the Prediction of Cancer Therapy	Raihan Rafique, S.M. Riazul Islam, Julhash U. Kazi	Comprehensive review of machine learning techniques for monotherapy and drug synergy prediction. Highlights emerging deep learning techniques like graph convolutional networks.	Limited clinical relevance due to lack of pharmacogenomic data and challenges in transitioning models to clinical practice(1-s2.0-S200103702100293…)

Machine Learning-Based Anti-Cancer Drug Response Prediction and Search for Predictor Genes Using Cancer Cell Line Gene Expression (2021)

•Authors: KexinQiu, JoongHo Lee, HanByeol Kim, Seokhyun Yoon, Keunsoo Kang

•Description: This study applied ElasticNet regression to predict drug responses based on gene expression data, achieving high prediction accuracy, particularly for drugs like AZD6244. It also identified predictor genes contributing to drug sensitivity. o Advantages: ElasticNet regression provides accurate drug response predictions by selecting key predictor genes, offering high performance for certain drugs. o Limitations: Gene pre-selection using Pearson correlation may introduce bias, affecting robustness. The study is limited to 12 drugs and 1,000 cell lines, restricting its generalizability.

Machine Learning in the Prediction of Cancer Therapy (2021)

•Authors: RaihanRafique, S.M. Riazul Islam, Jullash U. Kazi

• Description: This paper provides a comprehensive review of machine learning techniques applied to cancer therapy, focusing on monotherapy and drug synergy prediction. It highlights the growing importance of deep learning methods like graph convolutional networks in improving drug response predictions.

II. MODEL ARCHITECTURE

Based on the gene expression data of the cancer cell lines and the two types of response indicators, we used a machine learning algorithm to construct a drug response prediction model. We first pre-selected genes based on the p-value of Pearson correlation co efficient [11] and then used ElasticNet to predict drug response and to further select the predictor genes among the pre-selected ones. Specifically, we performed Elastic Net regression separately on the two response values, from which common predictor genes were identified. These common genes were used again to predict drug response hoping that the prediction performance is better than, or at least similar to, those obtained separately for the two response in dictators. To confirm biological significance of predictor genes, we provide heat map and gene

ISSN 2581-9429 IJARSCT

DOI: 10.48175/568

ontology analysis results. Fig. 1 shows the entire experimental workflow. Pre-processing Before processing the data, we took logarithm on IC50 and normalized the cell line gene expression data using the robust multichip average [12]

Feature selection based on Pearson correlation coefficient For some drugs, there are thousands of genes in the gene expression data, but not many genes have strong correlation with the drug responses. Therefore, it is very important to pre-select the relevant genes first. Although ElasticNet has capability for gene se lection, it is subject to data dependency and/or batch effect and, sometimes, it ignores genes that are really important to predict drug responses. In this paper, to overcome such problem, we used two-step gene selection, where we first used the Pearson correlation coefficient to pre-select genes and then applied ElasticNet to fine select the predictor genes. In particular, we used p-value of Pearson correlation coefficient between the drug response and the expression of each gene, with which genes with $p = 0.05$ or less were selected in the first feature selection. ElasticNet-based feature selection and drug response prediction ElasticNet [13] is a linear regression model trained with both $\ell 1$ and $\ell 2$ regularization. It is useful when there are so many features that are correlated with one another. In our data, the number of features (genes) is much larger than the number of samples and the prediction might be subject to overfit. Hence, to appropriately select genes and to suppress generalization error, we used Elastic Net to predict the drug response.

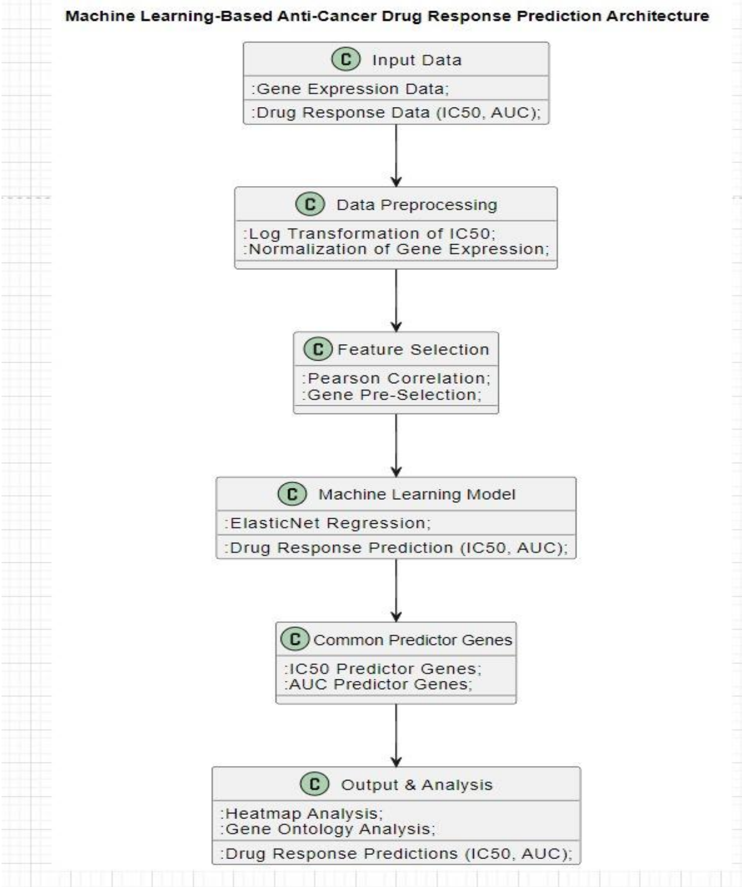


Fig.1System Architecture

The ElasticNet was selected based on the preliminary experiments where we compared Elastic Net with two well-known models, SVR [14] and Xgboost [15]. The former can be configured to a non-linear regressor by using various kernel functions and we used radial basis function kernel and the latter is an improved version of decision tree based gradient boosting algorithm. The two algorithms were shown to perform good for many applications, while, according to our preliminary experiments, they seem to have higher overfit than ElasticNet as the numbers of predictor genes that are common for the two response indicators were smaller than that for the ElasticNet. Fig. 2 summarizes the comparison for the 12 drugs in terms of Pearson correlation coefficients between the predicted IC50 and the measured ones.

DFT Diagram

DFD Elements:

- External Entities (Squares): These represent the actors who interact with the system (DataScientist, Researcher).
- Processes (Circles or Rounded Rectangles): These represent the actions or transformations performed on the data (Load & Preprocess, Gene Selection, Train Model, etc.).
- Data Stores (Open-Ended Rectangles): These represent where data is stored (Drug Response Data, Gene Expression Data, etc.).
- Data Flows (Arrows): These represent the movement of data between entities, processes, and data stores.

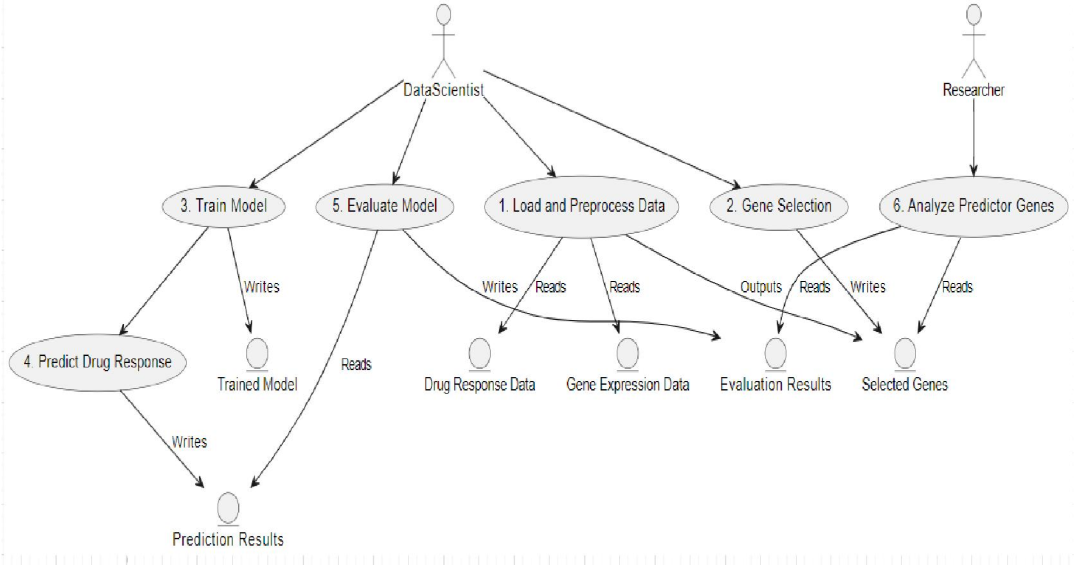


Fig.2DFD Diagram

Dataset

Gene expression microarray Gene expression microarray data for the NCI-60 cell lines were downloaded from the NCI DTP site (http://dtp.nci.nih.gov). Both Genelogic and Chiron NCI-60 datasets were originally generated using Affymetrix u133A/B microarrays and processed using MAS5. Probesets with a row-wise coefficient of variation (SD/mean) >0.06 were kept for further analysis. Probesets were further retained if they showed >0.2 Pearson's correlation coefficient across the 58 cell lines shared between Genelogic/Chiron datasets. Data for one glioma cell line (U251) that appeared both in the NCI-60 and 7 Glioma cell lines was dropped from the NCI-60 dataset before the generation of statistical models. Lastly, the arithmetic mean of matching probesets from Genelogic/Chiron datasets was taken, and these composite probesets were then z-normalized in a column-wise fashion for each cell line. Affymetrix u133 2.0+ Gene expression microarray data for 19 breast cancer cell lines (GSE3156) was downloaded from the NCBI Geo Gene Expression Database (http://www.ncbi.nlm.nih.gov/geo/). Probeset values were column-wise z normalized. Six cell lines from this dataset that also appeared in the NCI-60 were excluded from further analysis. Gene expression microarray data for seven Glioma cell lines (A172, LN229, T98G, U87, U118, U251, U373) were measured on the

Affymetrix u1332.0+platform.RNAextractionandanalysiswasperformedasdescribed previously (Li et al., 2008). Probeset values were processed using MAS5 and then z normalized in a column-wise fashion.

Drug sensitivity data IC50 is defined as the concentration of a compound required to produce 50% growth inhibition after 48h in a cell line relative to the control. NCI-60 IC50 data for a list of 40 federal drug administration (FDA)-approved oncology drugs were downloadedfrom the DTP web site and used as a training set. Values (previously −log10 transformed) were normalized over the [0,1] interval. If more than one experiment existed for each drug, the entry with the largest number of replicates was used. For the seven glioma cell lines in the test set, we measured percent growth inhibition relative to a control for the 40 drugs at five concentration points in triplicate: 50 µM, 5µM, 500nM, 50nMand5nM.CelllinesU87,U373and T98G were grown in modified eagle's medium (MEM) 10% fetal bovine serum (FBS), LN229 was grown in Dulbecco's modified Eagle's medium 10%FBS,U251wasgrowninRPMI1640 5%FBSandbothA172andU118 were grown in DMEM10%FBS.Cellswereseeded at 10000 cells/well in a 96-well plate in 150µM media/well. Viability assays were performed after 48h of initial seeding as described previously (Vichai and Kirtikara, 2006). IC50 calculations were performed by curve fitting of the data using the IC50 package for the R statistical computing environment. For the 19 Breast cancer cell lines in the test set, IC50 data for simvastatin and pepleomycin were downloaded from Supplementary Material associated with Mori et al. (2009).

Signature Generation

UsingRandomForestvariableimportancetocreateageneexpression signature for each drug Existing algorithms for producing gene expression signatures from drug response data compute univariate measures of differential gene expression between cell lines labeled sensitive or resistant. Two drawbacks exist with this approach: (i) definition of resistant and sensitive cell lines can be drug dependent, is arbitrarily defined and methods based on SD are only appropriate when IC50 values in the NCI-60 are normally distributed—which often not the case. (ii) Univariate differential gene expression cannot capture higher order gene–gene interactions that may be important for predicting drug response.Toimproveonthisapproach,weusedgeneexpressionsignaturegenerated by fitting a regression model between drug IC50 and basal gene expression using Random Forest. This machine-learning algorithm combines many single regression or classification trees into a large 'ensemble' to improve performance and robustness. Two methods are used to inject randomness into the ensemble—(i) Bootstrap aggregation: each tree is grown from a randomly selected subset of the training cases (defaults to 70% for regression). (ii) Random Subspace Method: the splitting variable for each node in an individual tree is selected from a random subset of the input variables (defaults to 30% for regression). Supplementary Figure S1provides a more detailed explanation of the Random Forest algorithm. Each tree in the forest is trained on a random subset of the cases and then tested on the remainingcases,called the out-of-bag cases (OOB).Tocompute variable importance, individual values of variables in the OOB are randomly switched with another variable in the OOB. Decrease in performance of each regression tree, measured by R2 for each variable after its value has been permuted, provides a measure of its importance in the regression model. To use the variable importance measure to create a signature, a Random Forest regression model using 25000 trees was trained on the normalized IC50 values for each drug, using all the gene expression probeset values from the composite Genelogic/Chiron dataset. Probesets for the signature were selected if they showed variable importance values >2 SD above the mean of all variable importance values for each particular drug. To improve performance for signature generation, we used the Simple Network of Workstations (SNOW) package for R to process many drugs simultaneously on a computational cluster.

DOI: 10.48175/568

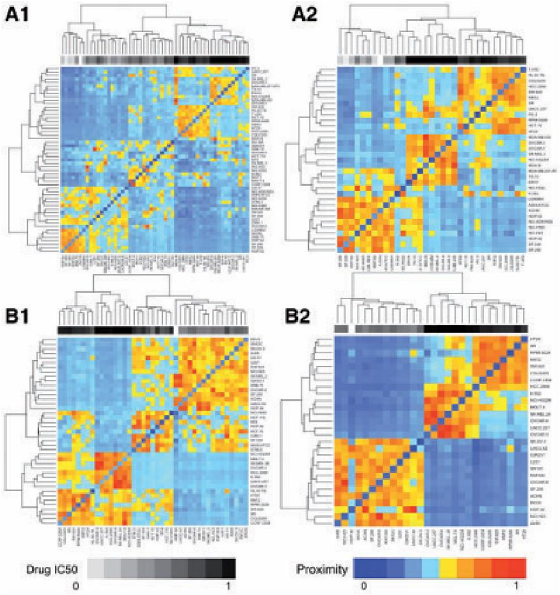


Fig. 3Pairwise proximity matrixes for pepleomycin and simvastatin. Proximity matrices

Fig. 3.Pairwise proximity matrixes for pepleomycin and simvastatin. Proximity matrices from Random Forest are defined as the number of instances in which two cases (cell lines) are assigned to the same terminal node of a tree, normalized over the [0,1] interval. Proximity between a case and itself is not a meaningful value so these instances on the diagonal are set to zero (A1) proximity matrix for pepleomycin before reduction of cell-lines by Equation (1). (A2) Proximity matrix for pepleomycin after removal of outlying cell lines. (B1) Proximity matrix for simvastatin. (B2) Proximity matrix for simvastatin after removal of outlying cell lines. After selecting core cell lines for each drug, the regression model was then built between the gene expression signature for these cell lines and the corresponding IC50 values for each drug using Random Forest with 10000 trees. The model for each drug was then applied to the drug gene expression signature in the test set.

IV. CONCLUSION

Ultimately, the findings of this study will contribute to the ongoing efforts in precision oncology, enabling healthcare professionals to tailor treatment regimens more effectively based on individual genetic profiles. By advancing our understanding of the genomic underpinnings of drug sensitivity, this research has the potential to improve patient outcomes and inform future therapeutic developments. Although the model proposed in this study shows good predictive performance for GDSC, there are still some limitations. First, the characteristic of cancer cell line may be quite different from the in vivo cancers and it should be verified whether this will be effective in clinical trial. Second, we perform drugs response prediction mainly based on gene expression data.

While, the response of drugs is not only related to gene expression levels, but also to structural variations such as gene mutations. Therefore, more study is required to utilize such information and integrate them into the model to improve the predictive power. Cancer is one of the leading causes of death worldwide. If one can find a new treatment by accurately predicting drug response, the probability of recovery will also be increased. Although there are still huddles to overcome in drug response prediction, advances in machine learning techniques will make it possible to introduce new ideas for drug response prediction that can provide accurate drug treatments and make it practical for clinicians and non-experts.

ISSN
2581-9429
IJARSCT

DOI: 10.48175/568

V. ACKNOWLEDGMENT

It gives us great pleasure in presenting the paper on “Anti-Cancer Drug Response Prediction using Gene Expression” We would like to take this opportunity to thank our guide, Prof.Alka Kumbhar, Professor, Department of Computer Engineering, Genba Sopanrao Moze College of Engineering, Balewadi, Pune, for giving us all the help and guidance we needed. We are grateful to her for her kind support, and valuable suggestions were very helpful.

REFERENCES

[1]. Rafique, R., Islamb, S. M. R., &Kazi, J. U. (2021). Machine Learning in the Prediction of Cancer Therapy. Computational and Structural Biotechnology Journal, 19, 4003-4017. https://doi.org/10.1016/j.csbj.2021.07.003

[2]. Wang, L., Li, X., Zhang, L., &Gao, Q. (2017). Improved Anticancer Drug Response Prediction in Cell Lines Using Matrix Factorization with Similarity Regularization. BMC Cancer, 17, 513. https://doi.org/10.1186/s12885-017-3500-5

[3]. Qiu, K., Lee, J. H., Kim, H. B., Yoon, S., & Kang, K. (2021). Machine Learning Based Anti-Cancer Drug Response Prediction and Search for Predictor Genes Using Cancer Cell Line Gene Expression. Genomics & Informatics, 19(1), e10. https://doi.org/10.5808/gi.20076

[4]. Geeleher, P., Cox, N. J., & Huang, R. S. (2014). PrISe: A method for predicting drug response in cancer cell lines. Biostatistics, 15(2), 374-385. https://doi.org/10.1093/biostatistics/kxt036

[5]. Menden, M. P., et al. (2013). Community assessment to advance a drug response predictive model. Nature Biotechnology, 31(12), 1251-1254. https://doi.org/10.1038/nbt.2726

[6]. Riddick, L. R., et al. (2011). Random Forest Regression for Drug Response Prediction in Cancer Cell Lines. Cancer Informatics, 10, 239-252. https://doi.org/10.1177/117693511100100012

[7]. Cortés-Ciriano, I., et al. (2020). Predicting drug response using molecular data. Journal of Molecular Biology, 432(2), 349-363. https://doi.org/10.1016/j.jmb.2019.09.00

ISSN 2581-9429 IJARSCT

DOI: 10.48175/568